COLABORACIÓN ESPECIAL

REGRESIÓN LOGÍSTICA NO CONDICIONADA Y TAMAÑO DE MUESTRA: UNA REVISIÓN BIBLIOGRÁFICA

Manuel Ortega Calvo (1) y Aurelio Cayuela Domínguez (2)

Centro de Salud Pilas, Sevilla

Unidad de Apoyo a la Investigación. Hospitales Universitarios Virgen del Rocío, Sevilla

RESUMEN

La regresión logística no condicionada es un método de predicción de riesgo muy útil en epidemiología. En este artículo revisamos las diferentes soluciones que han dado diversos autores sobre la interfase entre el cálculo del tamaño muestral y la utilización de la regresión logística. A partir del conocimiento de las primeras aportaciones, se revisan los fenómenos de regresión a la media y de la constricción predictiva, el diseño de una exposición ordinal con una salida binaria, el concepto de evento de interés por variable, las variables indicadoras, la fórmula clásica de Freeman, etc. Recogemos también algunas ideas escépticas sobre este tema.

Palabras clave: Regresión logística. Tamaño muestral. Diseño de estudios. Epidemiología.

ABSTRACT

Unconditioned Logistic Regression and Sample Size: A Reference Source Review

Unconditioned logistic regression is a highly useful risk prediction method in epidemiology. This article reviews the different solutions provided by different authors concerning the interface between the calculation of the sample size and the use of logistics regression. Based on the knowledge of the information initially provided, a review is made of the customized regression and predictive constriction phenomenon, the design of an ordinal exposition with a binary output, the event of interest per variable concept, the indicator variables, the classic Freeman equation, etc. Some skeptical ideas regarding this subject are also included.

Key words: Logistic regression. Sample size. Research Design. Epidemiology.

El modelado es la emoción que la mano experimenta en la caricia

Auguste Rodin

INTRODUCCIÓN

Uno de los elementos que más ha contribuido al avance de la epidemiología en los últimos años ha sido el desarrollo de determinados métodos de análisis como la regresión logística[1]. Mediante ella se pueden hacer cuantificaciones de riesgo en un determinado carácter biológico o no biológico (por ejemplo, el hábito tabáquico) permitiendo al investigador la creación de modelos uni o multivariantes que sean predictivos de fenómenos complejos. También ayuda a controlar el efecto de posibles variables confusoras y la interacción[2].

El modelo logístico aplicado a los estudios de seguimiento fue introducido por Cornfield en el año 1962[3] y posteriormente aplicado al análisis de los datos del estudio de Framingham[4]. La adaptación al contexto de estudios causales planteaba el problema de la estimación de los coeficientes, por lo que el uso de ordenadores era imprescindible.

El algoritmo de Walker-Duncan[5] para la obtención de los estimadores de máxima verosimilitud y los trabajos de Day y Kerrid-

Correspondencia:
Manuel Ortega Calvo.
Avda. de la Cruz del Campo 36.
Jardines de Villagracia. Bloque 1. 2.º A.
41005 Sevilla.

ge[6] y de Cox[7] vinieron a solucionar en parte este problema. Los estudios de bondad de ajuste descritos con posterioridad han aportado las técnicas de diagnóstico adecuadas[8-13]. La aplicación de los modelos logísticos en los estudios caso-control fue sugerida y justificada por Mantel[14] y por Siegel y Greenhouse[15]. Poco después se planteó la estimación de los coeficientes utilizando un argumento condicionado[16], lo cual permitía la aplicación en diseños pareados (regresión logística «condicionada»).

El objetivo de este trabajo es la revisión de las diferentes soluciones que distintos autores han dado al problema del cálculo del tamaño muestral para el caso en el que se aplique la regresión logística no condicionada en la modelación estadística de un estudio epidemiológico.

Primeras aportaciones

A pesar de que, como hemos visto, el método de análisis basado en el modelo logístico comenzó a existir científicamente a partir de la década de los sesenta[3-6], es interesante hacer notar que existe poca bibliografía específicamente dedicada a este modelo y al cálculo del tamaño de muestra hasta el año 1981[17]. Basada en una matriz de información para los parámetros estimados de una regresión logística múltiple y en una aproximación a ella mediante otra matriz para las covariables, Whittemore publicó una solución de tamaños de muestra en circunstancias uni y multivariadas para eventos raros[17]. En el año 1989 aparece otra publicación con objetivos similares[18]. Basado matemáticamente en el de Whittemore, Hsieh escribe de forma más comprensible, con unas tablas muy claras que contemplan las diferentes prevalencias asumidas en la enfermedad a estudiar y las odds ratio que se pretenden detectar (tabla 1), si la variable independiente fuera de tipo dicotómico y si se tratara de estudios caso-control apareados nos remite a otras fuentes[19-20]. Aunque no refiere nada acerca de las técnicas analíticas para el cumplimiento de normalidad uni o multivariante[21-22] si que recomienda que si se constata su falta en una o varias de las covariables se realice una transformación[23] antes de aceptar el tamaño de muestra[18].

Tabla 1

Tamaños de Muestra para Regresión Logistica Univariante

P \ r	*0,6*	*0,9*	*1,3*	*1,7*	*2,0*	*3,0*
0,01	2.334	56.741	9.076	2.158	1.237	480
0,05	518	12.251	1.970	481	285	139
0,10	291	6.689	1.181	271	166	96
0,16	206	4.604	749	192	121	80
0,25	155	3.352	549	145	94	71
0,40	121	2.518	416	114	76	64
0,50	110	2.240	372	103	70	62

Tomado de F.Y. Hsieh (18).
P = proporcion general de eventos; r = Odds Ratio
alfa = 5% y poder = 80%.

Casi diez años más tarde, en 1998, Hsieh y cols. han publicado un nuevo trabajo[24] en el que abordan el mismo problema pero con una metodología más clásica de formulación (cálculo de tamaño muestral para comparación de medias o para comparación de proporciones), no asumiendo que exista una probabilidad de respuesta baja en el modelo logístico[17-18].

Una cuestión fundamental: la regresión a la media

Entre los usos más frecuentes de la regresión en los estudios biomédicos está el de intentar describir cómo los valores de la variable dependiente están relacionados con el de la variable predictora o predictoras, el de intentar explicar cuáles son las predictoras de mayor interés y finalmente el intentar predecir, cuantificando su riesgo de aparición, los casos futuros[25].

El fenómeno de regresión a la media, desde que lo describió Francis Galton en el siglo XIX[25], implica que los valores nuevos de

la variable dependiente estarán más cercanos a la media aritmética general de lo que pudiera esperarse si se utiliza el método de los mínimos cuadrados ordinarios (regresión lineal, regresión polinomial, regresión exponencial) o el de la máxima verosimilitud (regresión logística). Este principio estadístico ocasiona una tendencia a la constricción *(shrinkage*) en la función predictiva de la regresión[25-26]. La constricción predictiva es mayor cuando existe una muestra pequeña o cuando se manejan muchas variables predictoras[25-26]. El término constricción fue utilizado por vez primera por Stein[27] en su trabajo sobre la estimación de la media normal multivariada.

Modelo con salida binaria y exposición ordinal

Un principio esencial en la determinación del tamaño de una muestra es que la aproximación utilizada se corresponda con los objetivos y el diseño de la investigación y con el tipo de análisis que se está planeando[28-29]. Phillips y Pocock[30] evaluaron las ideas primitivas de Whittemore[17] empleando datos de estudios prospectivos sobre enfermedad coronaria. Estos métodos están basados en la teoría de muestras grandes para estimación de la máxima verosimilitud en modelos lineales generalizados[31-32] o en la estimación de los mínimos cuadrados ponderados en modelos de tipo GSK[33-35].

Los modelos GSK se denominan así en honor a Grizzle, Starmer y Koch, los cuales los describieron en el año 1969 como una aproximación al análisis de datos cuando la variable de salida era medida en una escala nominal u ordinal[33]. Ellos contemplaron la posibilidad de valores perdidos de una forma aleatoria a lo largo de un estudio longitudinal[34].

Con todos estos antecedentes, Bull[28] expuso un método aplicable a datos sobre biología de los sarcomas óseos para el caso de la existencia de una variable de salida de tipo binario junto con una covariable ordinal. Se trataba de investigar el valor pronóstico de la expresión en el tejido sarcomatoso del gen de resistencia a drogas múltiples (mdr1)[28].

Para investigar el nivel de conservadurismo con respecto a la utilización de una escala de exposición más continua, se realizaron cálculos del tamaño de la muestra para la existencia de seis niveles de exposición, en lugar de los tres para los que al principio se había diseñado la investigación. Se esperaba que la utilización de seis niveles diera una aproximación más cercana a los requerimientos reales de tamaño de la la muestra basados en una medida de la exposición de tipo más continuo[36]. Sin embargo, utilizando tres niveles de exposición en vez de seis, tan sólo se producía un modesto aumento en el tamaño de la muestra de entre el 8% y el 12% dependiendo de la frecuencia de distribución de la exposición.

En síntesis, Bull expuso[28] una forma de estimar el tamaño de la muestra adecuado si se quiere detectar una tendencia lineal en el logaritmo del riesgo estimado de una respuesta binaria, basándose en contrastes de hipótesis. La expresiones para la varianza bajo la hipótesis alternativa podrían ser utilizadas también para la determinación de tamaños de muestra estimados dentro de unas tolerancias específicas[37], pero necesitarían de pequeños ajustes de muestra semejantes a los expuestos por Kupper y Hafner[38].

Otros Diseños

Flack y Eudey[39] desarrollaron un método para el cálculo del tamaño muestral en base al conocimiento previo de una matriz de datos que contenía información sobre la frecuencia de los factores y por lo tanto de la variabilidad del parámetro poblacional. Los tamaños de muestra «logísticos» están basados en la estimación de un parámetro de muestra pequeña utilizando un modelo que

incorpora una aproximación a la varianza de una muestra grande.

Los estudios en dos etapas buscan aumentar la eficiencia sin aumentar los costes. En ellos se mide la exposición y el resultado en una muestra grande (primera etapa) mientras que las covariables (estudio de confusión) tan sólo se analizan en una muestra más pequeña (segunda etapa)[40]. La solución del tamaño muestral en los diseños de dos etapas ha sido expuesta por varios autores[40-45].

El concepto de «evento de interés por variable»

En regresión logística hay que tener algunas precauciones cuando el número de covariables es elevado[46]. Freeman[47] sugirió que el número de sujetos para utililizarla sin problemas debía de ser superior a 10 * (k + 1), donde k expresa el número de covariables. Es decir, el tamaño muestral había de ser diez veces el número de parámetros a estimar más uno.

Por lo tanto si se introducen interacciones o variables indicadoras (dummy), el número de elementos muestrales debe de crecer de acuerdo con esta regla. Se ha sugerido asimismo que si una variable dicotómica (en especial si es la variable respuesta) no tiene al menos 10 casos en cada uno de sus valores posibles las estimaciones no son fiables[48].

Hemos de definir en términos de análisis multivariante varios tipos de error que pueden cometerse a la hora de la modelización[48]. El error tipo I tiene lugar cuando se «sobreoptimiza» o se «sobreajusta» un modelo; ocurre cuando se seleccionan muchas variables para el modelo final, algunas de las cuales pueden acarrear ciertas irregularidades. El error tipo II se produce cuando se «infraoptimiza» o se ajusta el modelo final por debajo del nivel deseado no incluyéndose variables relevantes en el mismo. El error de tipo III se produce al realizar una optimización o «ajuste paradójico», asignando una dirección incorrecta de asociación a una variable. Para evitar estos tipos de error se han descrito diversas estrategias. Harrell y cols.[49] enunciaron dentro de un marco teórico el criterio de un mínimo de 10-20 eventos por variable. En una simulación, Freedman y Pee[50] demostraron que el error tipo I aumentaba cuando la razón del número de variables con respecto al número de observaciones era mayor de 0,25, correspondientes a una tasa de eventos por variable inferior a 4.

Como el impacto del concepto «eventos o sucesos por variable» sobre los diversos métodos de análisis multivariante no es el mismo[51-52], Perduzzi y cols. realizaron una simulación tipo Monte Carlo para la medición de su efecto en los análisis de datos realizados mediante regresión logística[48]. Se utilizaron los datos de un ensayo clínico cardiológico con 673 personas, con un total de 252 muertes observadas[53]. Se seleccionaron siete variables pronósticas para su análisis, por lo que existía en la muestra original una tasa de eventos por variable de 252/7 = 36.

Para eventos o sucesos por variable menores de 10, los coeficientes de regresión se veían claramente sesgados tanto en sentido positivo como negativo y se observaba un aumento en las asociaciones paradójicas con significación en la dirección errónea (error tipo III). En análisis de simulación realizados con anterioridad, se habían observado problemas de exactitud y de precisión de las estimaciones, así como en la significación[54].

Con un enfoque parecido al que hemos descrito hasta ahora de «evento o suceso por variable», Irala y cols.[55] publicaron un trabajo explorando la posibilidad de qué ocurre cuando se crea una variable indicadora (dummy) con una frecuencia igual a cero en una de sus celdas.

Las variables indicadoras o dummy han de crearse en el entorno de la regresión logística para salvar el escollo conceptual que supone

la existencia de variables cualitativas o nominales. Si la variable nominal posee C categorías han de crearse [C-1] variables dicotómicas indicadoras que tan sólo contengan los valores 0 y 1 en un determinado orden.

Irala y cols.[55] enfrentaron una matriz de datos conteniendo algunas variables indicadoras con celdas de frecuencia igual a cero a ocho de los programas de análisis estadísticos más comunes para este tipo de estudios (BMDP, EGRET, JMP, SAS, SPSS, STATA, STATISTIX y SYSTAT) y observaron lo que pasaba. En todos se llegaba a resultados incongruentes que no debían ser difundidos.

Muestreo por Conglomerados

El muestreo por conglomerados *(clusters)* se define como aquel método de muestreo probabilístico en donde la unidad seleccionada es un grupo de individuos (por ejemplo todos los que viven en un bloque de pisos, una familia...), en lugar de un individuo en particular[56]. Hendricks y cols. publicaron un artículo[57] en el que comparaban la regresión logística con la utilización de ecuaciones estimativas generalizadas como método de cálculo del poder de una muestra para datos distribuidos en forma de conglomerados. El problema fundamental en este tipo de análisis es la correlación dentro del conglomerado (intracluster)[58-59].

Una visión escéptica del problema

Aunque lo dejó esbozado en el capítulo de tamaño muestral de su monografía sobre regresión logística[2], Silva lo ha reafirmado en su libro de cultura estadística en la investigación biomédica[60]. En términos generales es necesario suponer un error máximo que pueda ser aceptado y del que a veces no resulta fácil su identificación a priori. Quizás sea debido a que esta parte de la estadística necesite de una solución bayesiana pura que la libere de subjetividades[61-64].

Otro aspecto muy importante es la estimación de parámetros múltiples en el análisis y comentario final de los trabajos de investigación epidemiológica, cuando es frecuente que el cálculo del tamaño de la muestra se realice generalmente con una visión monovariante hacia un determinado parámetro. En los textos clásicos[65] este detalle no está muy bien resuelto.

Es muy importante el comentario que realiza Silva[60] sobre el carácter reductor de los tamaños muestrales en estudios con post-estratificacion, que son la mayoría, pues entonces se están analizando unos «n» menores a los calculados al principio del trabajo. Para nosotros, Silva

deja entrever cierto escepticismo en la solución al problema del cálculo del tamaño muestral[2,60].

CONCLUSIONES

En el ámbito epidemiológico la respuesta más común cuando se pregunta por el tamaño de la muestra en un estudio con regresión logística es la fórmula clásica de Freeman[47]: [n = 10 * (k + 1)] o lo que es lo mismo, en términos generales, el tamaño de muestra ha de ser unas diez veces el número de variables independientes a estimar más uno.

Para nosotros la idea más interesante de todas las expuestas en esta revisión bibliográfica es la de Perduzzi y cols. sobre suceso o evento de interés[48]. Llegan a resultados bastante parecidos a los de Freeman[47], en los que los cálculos pierden exactitud y precisión a medida que la proporción de eventos por variable baja de diez. Pero hay que tener cuidado, ya que este concepto es diferente al anterior. Freeman estimaba un total de diez elementos muestrales por cada variable a evaluar, mientras Perduzzi et al. hablan de «diez eventos de interés por variable» que pueden significar más de diez elementos[48].

Se ha estudiado también la repercusión que tiene un número bajo de eventos de inte-

rés por variable a la hora del sesgo en la selección automática *(stepwise)* de variables en un modelo múltiple, llegándose a la conclusión de que el sesgo es mayor a medida que disminuyen el número de eventos de interés[66]. Las técnicas de diagnóstico persiguen estudiar si los datos observados se adecuan al modelo propuesto[8-13,67]. La monografía de Sánchez-Cantalejo es muy explicativa en este y otros muchos sentidos[67].

Aunque la representatividad muestral siga siendo esencialmente intuitiva[68], quizás la utilización de algunas reglas metodológicas ayuden al investigador a decidir hasta dónde debe llegar en la recogida de datos.

Comentarios a originales

De forma totalmente aleatoria hemos elegido dos trabajos publicados en castellano para estudiar los tamaños muestrales utilizados. El primero de ellos es el de González-Clemente y cols.[69] realizado sobre un grupo de ancianos y tendente al estudio de la prevalencia de la hipovitaminosis D y de sus factores asociados. Se incluyeron al final del proceso selectivo un total de 127 sujetos (47 varones y 80 mujeres). Después del análisis bivariante de los factores relacionados con el problema, se obtuvieron un total de diez variables significativas ($p < 0,05$) que se introdujeron en el modelo logístico. La presencia/ausencia de déficit de vitamina D fue considerada como variable dependiente. De esta modelación salieron cinco variables en el ajuste final [edad, OR = 1,17, exposición solar, OR = 0,32, albuminemia, OR = 0,05, talla, OR = 0,01 y fosforemia, OR = 0,31]. Contemplado desde el concepto de evento de interés por variable[48] esta modelación contiene después del ajuste cinco variables asociadas. De los 127 individuos de la muestra, 44 presentaban valores compatibles con hipovitaminosis D (evento de interés). Por lo tanto, su proporción de eventos de interés por variable es de 44/5 = 8,8. Es un valor cercano a diez y aceptable. Entre los puristas de este método cabría discutir si la proporción debería ser aplicada al número total de variables introducidas a priori (en este caso diez). Nosotros entendemos que es aceptable el cálculo sobre el modelo final ajustado.

Según el trabajo clásico de Hsieh[18] (tabla 1) un diseño con 127 elementos muestrales está sobre una capacidad de detección de OR de 1,7 y una proporción general de eventos de entre 0,25 y 0,40 que creemos también adecuadas al diseño y a los resultados obtenidos[69].

El otro original que queremos comentar es el de García Lirola y cols.[70] sobre la adopción de nuevos medicamentos en un grupo de médicos prescriptores en atención primaria. Se recogió una muestra de 74 profesionales de medicina general con más de tres años de ejercicio. A partir del rastreo de prescripción de medicamentos considerados como novedad terapéutica y después de aplicar una escala mixta cuantitativo-temporal[70], los autores identificaron un total de 25 médicos innovadores en la muestra. En el modelo logístico explicativo de las características del médico innovador (variable dependiente: innovador/no innovador) introdujeron un total de ocho factores de los que arrojaron resultados significativos cuatro, casi todos ellos binarios[70]. Estos datos arrojan una proporción de eventos de interés por variable[48] de 25/4 = 6,25 que estimamos baja.

Según Hsieh[18] (tabla 1) un tamaño muestral de 74 sujetos rige una proporción de eventos de entre 0,40 y 0,50 y una capacidad de detección de OR de 2,0. Los autores[70] reconocen en la discusión que han manejado una muestra pequeña y en las conclusiones dan relevancia tan sólo a dos variables del modelo logístico.

BIBLIOGRAFÍA

1. Sanz Aguado M.A. La epidemiología y la estadística. En: Sánchez-Cantalejo Ramírez E.(editor). La epidemiología y la estadística. V Encuentro Marcelino Pascua. Granada.1995. Ponencias.

Granada: Publicaciones de la Escuela Andaluza de Salud Pública; 1996. p. 35-44.

2. Silva Ayçaguer LC: Excursión a la regresión logística en ciencias de la salud. Madrid: Díaz de Santos; 1995.

3. Cornfield J. Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. In: Federation Proceeding; 1962; 21. p. 58-61.

4. Truett J, Cornfield J, Kannel W. A multivariate analysis of the risk of coronary heart disease in Framingham. JChron Dis 1967; 20: 511-524.

5. Walker SH, Duncan DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 1967; S4: 167-179.

6. Day NE, Kerridge DF. A general maximum likelihood discriminant. Biometrics 1967; 23: 313-323.

7. Cox DR. Analysis of binary data. London: Methuen & Co; 1970.

8. Hosmer DW, Lemeshow S. A goodness-of-fit test for the multiple logistic regression model. Communications in Statistics 1980; A10: 1043-1069.

9. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley; 1989.

10. Le Cessie S, van Houwelingen JC. A goodness-of-fit test for binary data based on smoothing residuals. Biometrics 1991; 47: 1267-1282.

11. Le Cessie S, van Houwelingen JC. Testing the fit of a regression model via score tests in random effect models. Biometrics 1995; 51: 600-614.

12. Tsiatis AA. A note on a goodness-of-fit test for the logistic regression model. Biometrika 1980; 67: 250-251.

13. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Statistics in Medicine 1997; 16: 965-980.

14. Mantel N. Synthetic retrospective studies and related topics. Biometrics 1973; 29: 479-486.

15. Seigel DG, Greenhouse SW. Validity in estimating relative risk in case-control studies. J Chron Dis 1973; 26: 219-226.

16. Prentice RL, Breslow NE. Retrospective studies and failure time models. Biometrika 1978; 65: 153-158.

17. Whittemore A. Sample size for logistic regression with small response probability. J Am Statistical Association 1981; 76: 27-32.

18. Hsieh F. Sample size tables for logistic regression. Statistics in Medicine 1989; 8: 795-802.

19. Fleiss J. Statistical methods for rates and proportions. New York: Wiley; 1981.

20. Dupont WD. Power calculations for matched case-control studies. Biometrics 1988; 44: 1157-68.

21. Shapiro SS, Wilk MB. An analysis of variance test for normality (complete samples). Biometrika 1965; 52: 591-611.

22. Royston P. Estimating departure from normality. Statistics in Medicine 1991; 10: 1283-1293.

23. Sokal RR, Rohlf FJ. San Francisco: Biometry Freeman; 1969.

24. Hsieh FY, Bloch DA, Larsen M. A simple method of sample size calculation for linear and logistic regression. Statistics in Medicine 1998; 17: 1623-34.

25. Copas JB. Using regression models for prediction: shrinkage and regression to the mean. Statistical Methods in Medical Research 1997; 6: 167-183.

26. Copas JB, Jones M C. On the robustness of shrinkage predictors in regression to differences between past and future data. J Royal Statistical Society B 1986; 48: 223-237.

27. Stein C. Inadmissibility of the usual estimator of the mean of a multivariate normal distribution. Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability; 1956; 1: 197-206.

28. Bull SB. Sample size and power determination for a binary outcome and an ordinal exposure when logistic regression analysis is planned. Am J Epidemiol 1993; 137: 676-84.

29. Breslow NE, Day N E. Statistical methods in cancer research. Vol I. The analysis of case-control studies.(IARC scientific publications no. 32). Lyon: IARC; 1980.

30. Phillips AN, Pocock SJ. Sample size requeriments for prospective studies with examples for coronary heart disease. J Clin Epidemiol 1989; 42: 639-648.

31. Wilson SR, Gordon I. Calculating sample sizes in the presence of confounding variables. Appl Statist 1986; 35: 207-13.

32. Self SG, Mauritsen R H. Power/sample sizes calculations for generalized linear models. Biometrics 1988; 44: 79-86.

33. Rochon J. The application of the GSK method to the determination of minimun sample sizes. Biometrics 1989; 45: 193-205.

34. Park T, Davis CS. A test of the missing data mechanism for repeated categorical data. Biometrics 1993; 49: 631-8.

35. Hardison CD, Quade D, Langston RD. Nine functions for probability distributions. In: SAS Institute Inc. SUGI supplemental library user's guide. Version 5 edition. Cary, NC: SAS Institute Inc.; 1986. p. 385.

36. Howe GR, Chiarelli A M. Methodological issues in cohort studies II Power calculations. Int J Epidemiol 1988; 17: 464-8.

37. Greenland S. On sample-size and power calculations for studies using confidence intervals. Am J Epidemiol 1988; 128: 231-7.

38. Kupper LL, Hafner KB. How appropiate are popular sample size formulas? The American Statician 1989; 43: 101-5.

39. Flack VF, Eudey TL. Sample size determinations using logistic regression with pilot data. Statistics in Medicine 1993; 12: 1079-1084.

40. Cain KC, Breslow NE. Logistic regression analysis and efficient design for two-stage studies. Am J Epidemiol 1988; 128: 1198-1206.

41. White JE. A two-stage design for the study of the relationship between a rare exposure and a rare disease. Am J Epidemiol 1982; 115: 119-128.

42. Walker AW. Anamorphic analysis: sampling and estimation for covariate effects when both exposure and disease are known. Biometrics 1982; 38: 1025-1032.

43. Zhao LP, Lipsitz S. Designs and analysis of two-stages studies. Stat Med 1992; 11: 769-782.

44. Reilly M. Optimal sampling strategies for two-stage studies. Am J Epidemiol 1996; 143: 92-100.

45. Schaubel D, Hanley J, Collet JP, Boivin JF, Sharpe C, Morrison HI, Mao Y. Two-stage sampling for etiologic studies. Sample size and power. Am J Epidemiol 1997; 146: 450-458.

46. Álvarez Cáceres R. El método científico en las ciencias de la salud. Las bases de la investigación biomédica. Madrid: Díaz de Santos; 1996. p. 29-55.

47. Freeman DH. Applied categorical data analysis. New York: Marcel Dekker Inc; 1987.

48. Perduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49: 1373-9.

49. Harrell F, Lee KL, Matchar DB, Reichert TA. Regression models for prognostic prediction: Advantages, problems and suggested solutions. Cancer Treatment Reports 1985; 69: 1071-7.

50. Freedman LS, Pee D. Return to a note on screening regression equations. Am Statistician 1989; 43: 279-282.

51. Concato J, Perduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals and general strategy. J Clin Epidemiol 1995; 48: 1495-1501.

52. Concato J, Perduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995; 48: 1503-10.

53. Perduzzi P, Detre K, Gage A. Veterans administration cooperative study of medical versus surgical treatment for stable angina. Progress report: Section 2 - Design and baseline characteristics. Prog Card Dis 1985; 28: 235-243.

54. Lilienfeld D, Pyne DA. The logistic analysis of epidemiologic prospective studies:Investigations by simulation. Stat Med 1984; 3: 15-26.

55. De Irala J, Fernández-Crehuet Navajas R., Serrano del Castillo A. Intervalos de confianza anormalmente amplios en regresión logística: interpretación de resultados de programas estadísticos. Rev Panam Salud Pública. 1997; 3: 230-233.

56. Donner A, Klar N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. Am J Epidemiol 1994; 140: 279-289.

57. Hendricks SA, Wassell JT, Collins JW, Sedlak SL. Power determination for geographically clustered data using generalized estimating equations. Statistics in Medicine. 1996; 15: 1951-1960.

58. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 78: 13-22.

59. Zeger S.L.,Liang K.Y.: Longitudinal data analysis for discrete and continuous outcomes.Biometrics. 1986; 42: 121-130.

60. Silva Ayçaguer LC. El enigma del tamaño muestral. En: Silva Ayçaguer, L. C.: Cultura estadística e Investigación científica. Ed. Díaz de Santos. S.A. Madrid. 1997. p. 285-305.

61. Brunier HC, Whitehead J. Sample sizes for phase II clinical trials derived from Bayesian decision theory. Statistics in Medicine. 1994; 13: 2493-2502.

62. Marchiset-Leca D, Leca FR, Galeani A, Noble A, Iliadis A. A limited sampling strategy for the study of pirarubicin phamacokinetics in humans. Cancer Chemoter Pharmacol 1995; 36: 233-238.

63. Bacallao J. La perspectiva exploratorio-confirmatoria en las aplicaciones biomédicas de la estadística: dos diálogos (I). Bayesianismo frente a frecuencialismo: sus respectivas implicaciones prácticas en relación con el análisis de datos. Med Clín (Barc) 1996; 107: 467-471.

64. Bacallao J. La perspectiva exploratorio-confirmatoria en las aplicaciones biomédicas de la estadística: dos diálogos (y II).Consideraciones críticas acerca de las pruebas de significación. Med Clín (Barc) 1996; 107: 539-543.

65. Lemeshow S, Hosmer Jr, DW, Klar J, Lwanga SK: Adequacy of sample size in health studies. New York: Wiley; 1990.

66. Steyerberg EW, Eijkemans MJC, Habemma DF: Stepwise selection in small data sets: A simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999; 52: 10: 935-42.

67. Sánchez-Cantalejo Ramírez E. Regresión Logística en Salud Pública. Granada: Escuela Andaluza de Salud Pública. Monografías. 2000; núm. 26.

68. Silva Ayçaguer LC. Diseño razonado de muestras y captación de datos para la investigación sanitaria. Madrid: Díaz de Santos; 2000. p. 19-20.

69. González-Clemente JM, Martínez-Osaba MJ, Miñarro A, et al. Hipovitaminosis D: alta prevalencia en ancianos de Barcelona atendidos ambulatoriamente. Factores asociados. Med Clín (Barc). 1999; 113: 641-645.

70. García Lirola MA, Cabeza Barrera J, Rodríguez Espejo M, et al. Adopción de los nuevos medicamentos por los médicos prescriptores. El médico innovador. Aten Primaria 2000; 25: 22-28.